Curation & Preservation of Crystallography Data

Chemistry in the Digital Age: A Workshop Connecting Research and Education

Department of Chemistry, Penn State University, PA, US
June 11-12th 2009

Manjula Patel
UKOLN, University of Bath, UK

Digital Curation & Preservation

- Maintaining and adding value to a trusted body of digital information for current and future use
- Encompasses the active management of data throughout the information lifecycle
- Requires a commitment to undertake duties of stewardship
- Influenced by a complex array of factors including social, political, cultural, organizational, financial, legal and technical issues
- Preservation is a pre-requisite to curation
- Curation and preservation are continuous processes

"Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate"

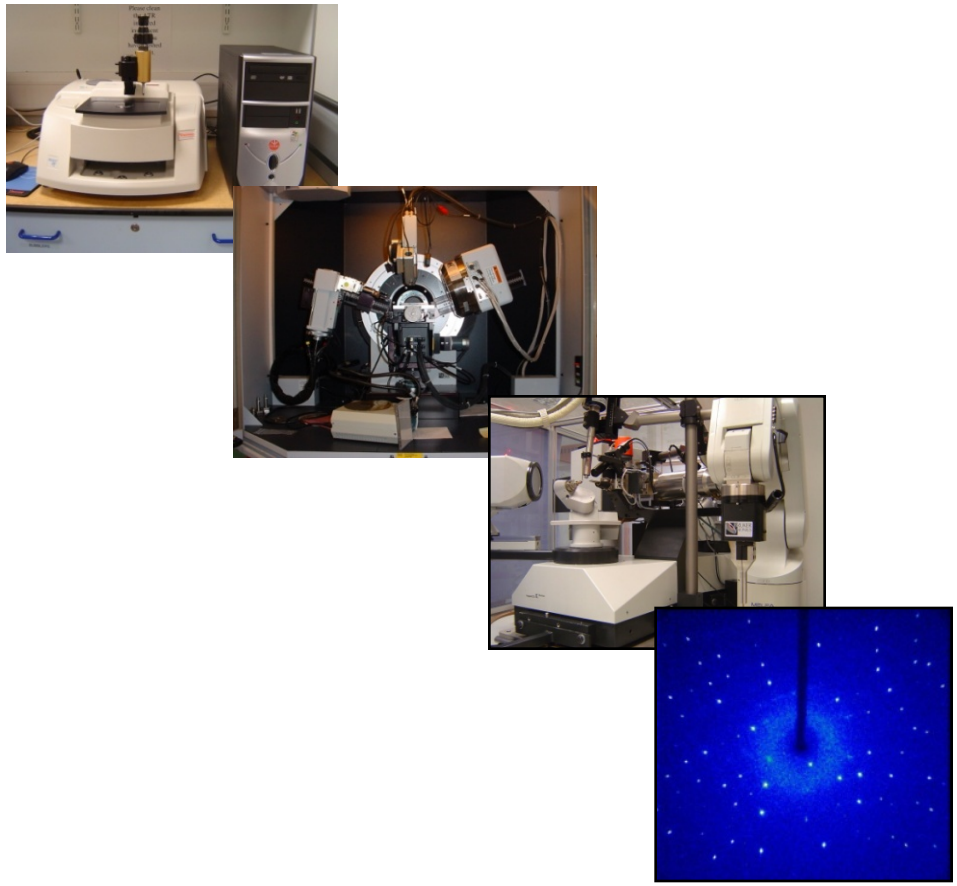
Clifford Lynch, Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age, ARL, No. 226, February 2003, pp1-7.

Curation & Preservation: Motivation

- Verification and validation of research results
- Scholarly knowledge lifecycle based on continuous use and reuse of data –derivative science
- Curated data has the potential to be re-purposed and generate new scientific results
- Recapturing and reproducing some data is difficult or impossible e.g. observational and environmental data
- Legal obligations –data retained over certain periods of time
- Research funding bodies
  - maximise reuse, cross-reference and dataset integration
  - ensure valuable datasets stored securely and remain accessible to future researchers
- Many workflows nowadays are almost completely digital

Crystallography –The Science

- Sub-discipline of chemistry
- Concerned with determining the structure of a molecule and its 3D orientation with respect to other molecules in a crystal
- Analysis of diffraction patterns obtained from X-ray scattering experiments
- Focus on laboratory based experimental technique of chemical crystallography undertaken at the EPSRC National Crystallography Service (NCS), UK



Images from Simon Coles (NCS), 2006

Community & Current Practice (1)

- Convention is to share results data, access to raw data is rare
- Crystallography Information File (CIF) is a de facto exchange standard
  - Maintained by International Union of Crystallography (IUCr)
- Heterogeneity in instrumentation and associated software
- Established system for publishing crystallographic data in UK
  - Cambridge Crystallographic Data Centre (CCDC)
- Other major databanks
  - Germany (inorganic molecule database); Canada (metals database); US (Protein Data Bank -PDB)
- Publishing datasets
  - Alongside journal articles through publisher mandates
  - Researchers often wish to retain exclusive use of their data

Community & Current Practice (2)

- "*To vet experiments, correct errors, or find new breakthroughs, scientists desperately need better ways to store and retrieve research data*"

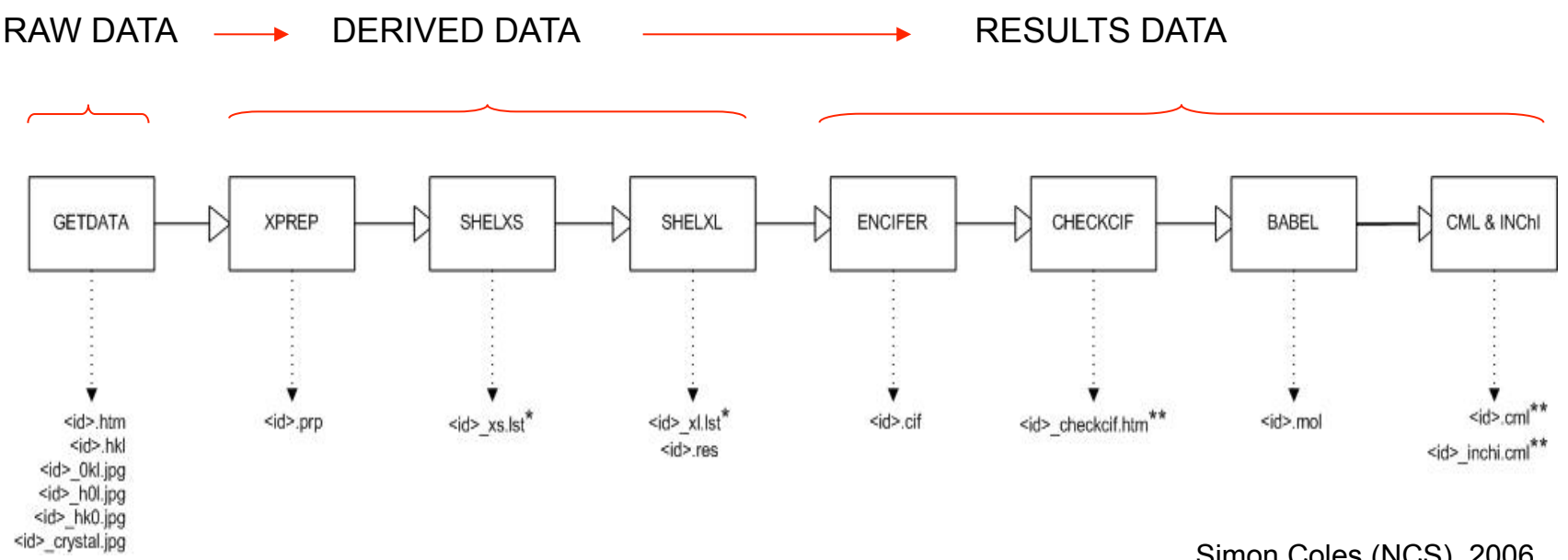
  Scott Carlson, Lost in a Sea of Science Data,
  The Chronicle of Higher Education, June 2006

  - Sometimes CIF retained but raw data discarded
  - Data often stored on DVDs or laptops
  - Distributed, local storage -shortage of local curation expertise
  - Quality of metadata for datasets is variable
- Open access
  - eCrystals Federation Project; CrystalEye; ReciprocalNet (US, Australia, UK); Crystallography Open Database (COD)
  - Chemistry Central (open access publisher)
- Open science
  - Open Notebook; MyExperiment; blogs

Building the eCrystals Repository

- eBank-UK Project
  - JISC funded; three phases Sept. 2003-June 2007
  - UKOLN (lead), EPSRC NCS, University of Manchester
- Phenomenal growth in amount of data generated from experiments
  - 40 years ago a PhD student would determine 2-3 structures for a thesis; this can now be easily achieved in a single day
- Only a small proportion is widely and easily accessible
  - Estimated that < 50% of crystal structures are published
  - Current data publication process is a bottleneck
- eCrystals data repository
  - Open access and rapid dissemination of derived and results data from crystallography experiments
  - Repository platform: ePrints.org software V3
  - Supported by learned society (IUCr) and subject repository (CCDC)
- Linking research data to publications and scholarly communication
- Metadata harvesting and aggregation (OAI-PMH)

EPSRC NCS Crystal Structure Determination Workflow

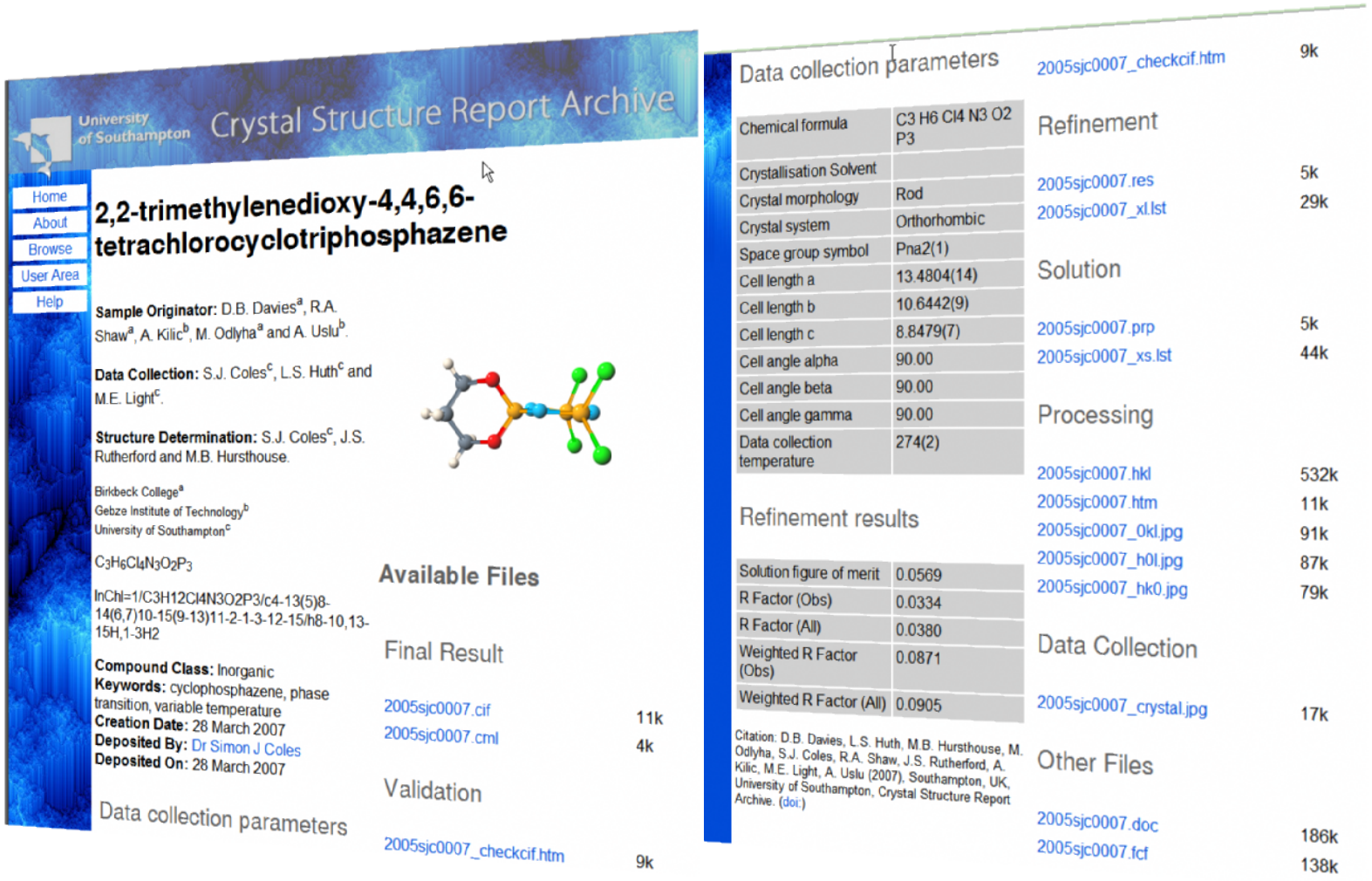


Simon Coles (NCS), 2006

- Initialisation: mount new sample
- Collection: collect data
- Processing: process and correct images
- Solution: solve structures

- Refinement: refine structure
- CIF: produce Crystallographic Information File
- Validation: chemical & crystallographic checks
- Report: generate Crystal Structure Report
- CML, INChI

eCrystals Data Repository: Example Crystal Structure Report



University of Southampton Crystal Structure Report Archive

Home
About
Browse
User Area
Help

2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: D.B. Davies[a], R.A. Shaw[a], A. Kilic[b], M. Odlyha[a] and A. Uslu[b].

Data Collection: S.J. Coles[c], L.S. Huth[c] and M.E. Light[c].

Structure Determination: S.J. Coles[c], J.S. Rutherford and M.B. Hursthouse.

Birkbeck College[a]
Gebze Institute of Technology[b]
University of Southampton[c]

$C_3H_6Cl_4N_3O_2P_3$

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/h8-10,13-15H,1-3H2

Compound Class: Inorganic
Keywords: cyclophosphazene, phase transition, variable temperature
Creation Date: 28 March 2007
Deposited By: Dr Simon J Coles
Deposited On: 28 March 2007

Data collection parameters

Available Files

Final Result

2005sjc0007.cif 11k
2005sjc0007.cml 4k

Validation

2005sjc0007_checkcif.htm 9k

Data collection parameters

Chemical formula	C3 H6 Cl4 N3 O2 P3
Crystallisation Solvent	
Crystal morphology	Rod
Crystal system	Orthorhombic
Space group symbol	Pna2(1)
Cell length a	13.4804(14)
Cell length b	10.6442(9)
Cell length c	8.8479(7)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	274(2)

Refinement results

Solution figure of merit	0.0569
R Factor (Obs)	0.0334
R Factor (All)	0.0380
Weighted R Factor (Obs)	0.0871
Weighted R Factor (All)	0.0905

Citation: D.B. Davies, L.S. Huth, M.B. Hursthouse, M. Odlyha, S.J. Coles, R.A. Shaw, J.S. Rutherford, A. Kilic, M.E. Light, A. Uslu (2007). Southampton, UK, University of Southampton, Crystal Structure Report Archive. (doi:)

2005sjc0007_checkcif.htm 9k

Refinement

2005sjc0007.res 5k
2005sjc0007_xl.lst 29k

Solution

2005sjc0007.prp 5k
2005sjc0007_xs.lst 44k

Processing

2005sjc0007.hkl 532k
2005sjc0007.htm 11k
2005sjc0007_0kl.jpg 91k
2005sjc0007_h0l.jpg 87k
2005sjc0007_hk0.jpg 79k

Data Collection

2005sjc0007_crystal.jpg 17k

Other Files

2005sjc0007.doc 186k
2005sjc0007.fcf 138k

Scoping Study Recommendations (1)

A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed Federation, M. Patel & S. Coles, Sept. 2007, eBank-UK Phase 3

- Develop a preservation and curation strategy and formal policies to indicate levels of service (e.g. deposit, ingest, validation, dissemination)
- Promote community-supported sustainability plan
- Self-assessment using DRAMBORA toolkit
  - Implement regular audits e.g. annually
  - Produce documentary evidence of compliance
- Maintenance and open access of critical file formats and software
  - Crystallography Information File (CIF)
  - Work-up software e.g. XPREP; SHELX{S,L}; ENCIFER; checkCIF; BABEL
  - Advocate export of raw data from instrumentation as IMG CIF

Scoping Study Recommendations (2)

- Capture relevant Representation Information
- Capture preservation metadata (e.g. versioning; provenance)
  - OAIS Preservation Description Information
  - PREMIS Data Dictionary
  - Extend or augment eBank-UK Metadata Application Profile
- Obtain consensus on Metadata Application Profile
- Seek to automate metadata generation, extraction and maintenance

eCrystals Federation Project

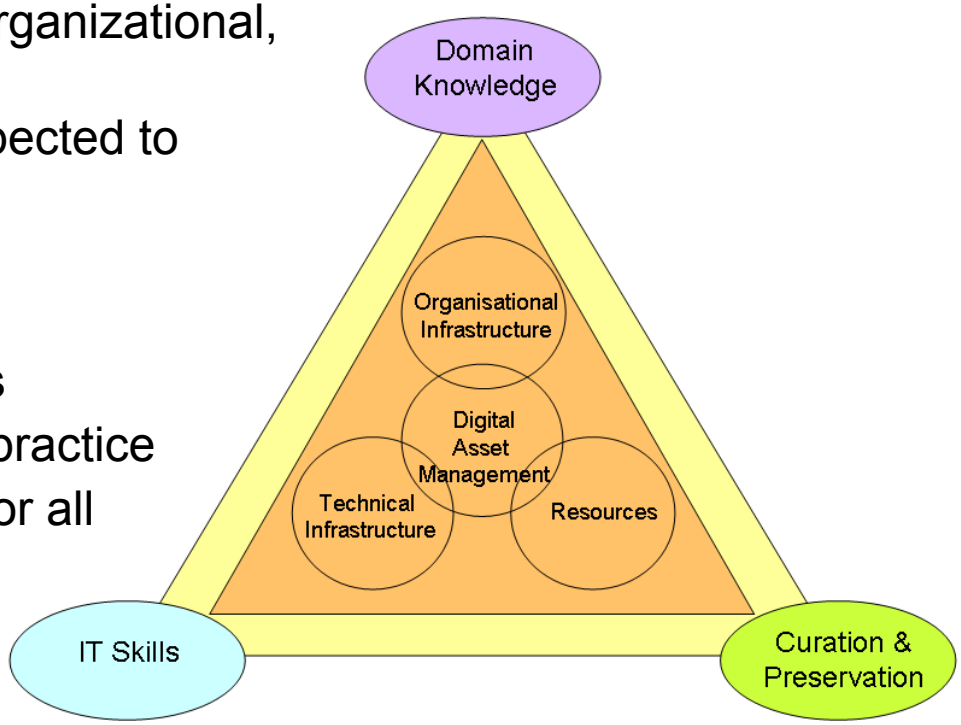
- Nov 2007–Mar 2009 (builds on eBank-UK Phase 3 results)
- Led by UK NCS with core partners at UKOLN (University of Bath), the Digital Curation Centre and the Unilever Centre (University of Cambridge) –14 supporting partners.
- Enhance management of crystallography data at institution level (data generated in departments, laboratories and by individual researchers or practitioners)
- Approaches to preservation and curation of scientific data in institutional repositories
  - Preservation Planning
  - OAIS Representation Information
  - Preservation Metadata
- Harmonise federation metadata application profile
- Investigate aggregation issues arising from harvesting metadata from federation repositories
- Enable interoperation of repositories with international subject archives (e.g.CCDC) and other third party harvesters

Preservation Planning

Preservation Planning

Preservation Planning for Crystallography Data,
M. Patel, June 2009, eCrystals Federation Project

- Commitments are influenced by a complex array of factors: social, cultural, political, organizational, financial, legal and technical
- Increasingly, repository manager expected to have skills in:
  - Information Technology
  - Research data domain
  - Curation and preservation issues
- Considerable diversity in laboratory practice
- No single set of guidelines suitable for all repositories



Components of Preservation Planning (1)

- Analyse Data & Associated Workflows
  - Understanding of the file formats as well as the inter-relationships between processing software and data files
  - Processes and workflows in crystallography labs differ considerably
- Evaluate Preservation Requirements
  - Life cycle of the data; how it is used and over what periods of time; user community
- Define a Preservation Policy
  - Access and reuse constraints
  - Deposition rules; quality control; copyright policy
  - Retention period; preservation strategy; withdrawal policy; version control

Components of Preservation Planning (2)

- Formulate a Preservation Strategy
  - Technological obsolescence of hardware, software and file formats
    - Raw data – images (JPEG); proprietary formats (.kcd)
    - Derived data – processed data (.hkl, .prp, .res, .lst etc.)
    - Results data - crystal structures (.cif, .cml, .mol)
    - Work-up software e.g. XPREP; SHELX{S,L}; ENCIFER; checkCIF; BABEL
  - Effective use of digital data dependent on:
    - complex configurations of hardware and software
    - semantics (bit-preservation vs. functional preservation)
  - Refresh, migrate, emulate, encapsulation (Representation Information)
  - Security, integrity, authenticity of data
  - Significant properties or characteristics
  - Consider options for out-sourcing specific functions
- Record Preservation Metadata

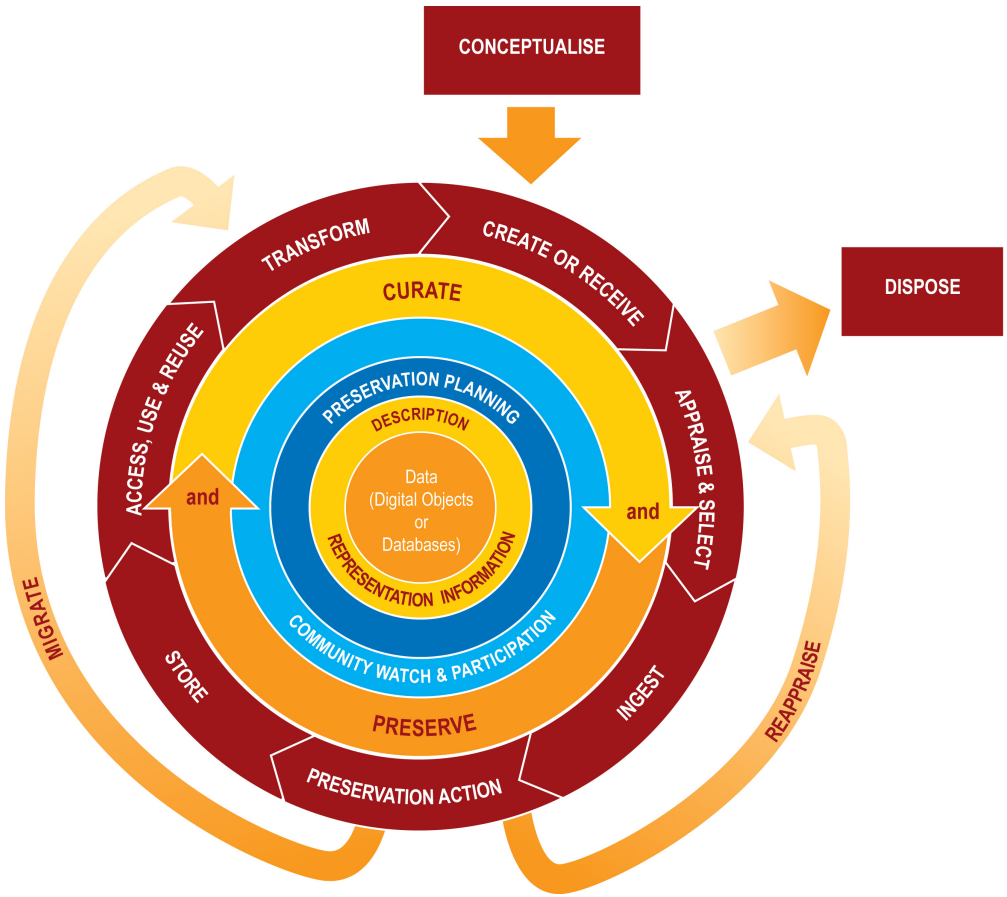
Components of Preservation Planning (3)

- Model Costs
  - Start-up, on-going and contingency
  - Repository's policy with regard to long-term management of data
  - Nature of the data; what and how much to preserve
  - Particular characteristics that need to be retained
  - Specific choices with respect to preservation strategy
  - Monitoring standards, technology, file formats, developments in user community
  - Staff training
- Plan for Sustainability
  - Gaining trust of data depositors and users
  - Business model; funding; repository and community levels
  - Forward planning for transitions in data stewardship
  - Interoperability (e.g. C/LOCKSS; OAI-ORE; OAI-PMH)
  - Community collaboration (e.g. Representation Information; metadata)

Components of Preservation Planning (4)

- Regular Evaluation and/or Self-Assessment
  - Verify periodically proper functioning of records, management procedures and systems
  - Validate integrity, authenticity and reliability of data
  - Audit and certification instruments
  - Revisit and revise preservation policies, strategies and plans

DCC Curation Lifecycle Model



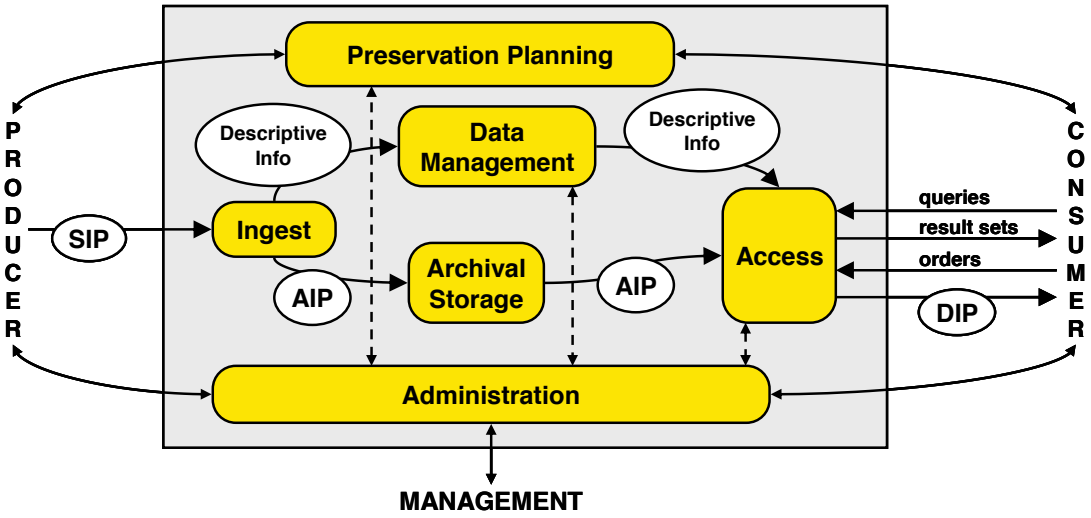
Digital curation necessitates a lifecycle approach

- how data is created will impact on how it can be managed and preserved in the future
- digital data is vulnerable throughout its entire lifetime

Full lifecycle stages (*Description & Representation Information; Preservation Planning; Community Watch and Participation; Curate & Preserve*)
Sequential actions (*Conceptualise; Create or Receive; Appraise and Select; Ingest; Preservation Action; Store; Access, Use & Reuse; Transform)*
Occasional actions (*Dispose; Reappraise; Migrate)*

Open Archival Information System (OAIS) Reference Model

- An OAIS is “An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community" (OAIS 1.7.2)
- Development led by the Consultative Committee for Space Data Systems (CCSDS)
- Adopted as ISO 14721:2003
- Reference model, not a blueprint for implementation
- Establishes a common framework of terms and concepts



- Identifies the basic functions of an OAIS:

  Ingest; Archival storage; Data Management; Administration; Preservation Planning; Access

OAIS: Mandatory Responsibilities

Covers a wide range of issues relating to the whole operating environment of a repository:

- Negotiating and accepting information
- Obtaining sufficient control of the information to ensure long-term preservation
- Determining the "designated community"
- Ensuring that information is "independently understandable"
- Following documented policies and procedures
- Making the preserved information available

Trust & Data Stewardship

Criteria for long-term repositories, meeting hosted by Center for Research Libraries, Jan 2007

1. Commits to continuing maintenance of digital objects
2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfil its commitment
3. Acquires and maintains requisite contractual and legal rights
4. Has an effective and efficient policy framework
5. Acquires and ingests digital objects based on stated criteria
6. Ensures integrity, authenticity and usability of digital objects
7. Creates and maintains metadata about actions taken on digital objects during preservation
8. Fulfils requisite dissemination requirements
9. Has a strategic program for preservation planning and action
10. Adequate technical infrastructure

Audit & Certification Instruments (1)

- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Feb 2007 (84 criteria):
  - Organisational infrastructure
  - Digital object management
  - Technologies, technical infrastructure and security
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), March 2007. A self-audit produces a composite risk score for each of eight functional classes, grouped into two types:
  - Organisational: *acquisition and ingest, storage and preservation, metadata management, access and dissemination*
  - Support: *organisation and management, staffing, financial management, technological solutions and security*

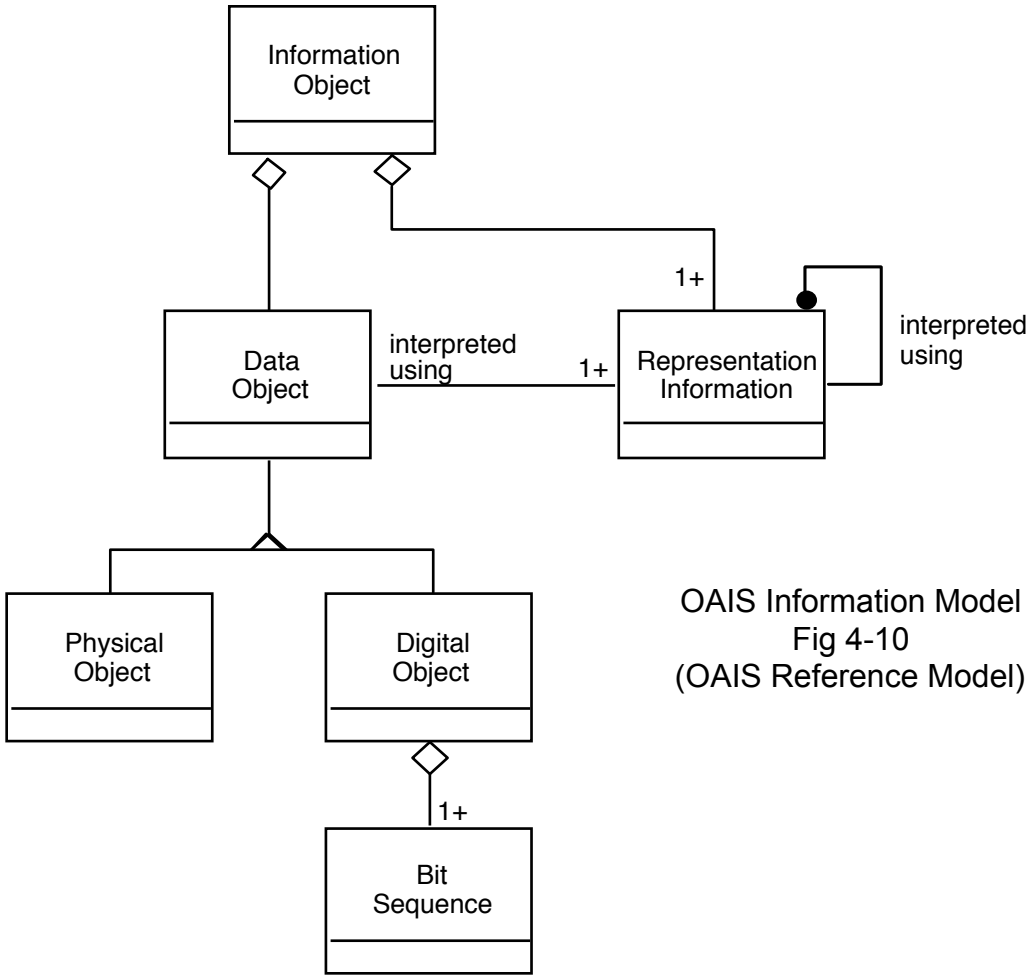
Audit & Certification Instruments (2)

- Planning Tool for Trusted Electronic Repositories (PLATTER), April 2006, EU PLANETS Project
  - Compliments other audit and certification tools
  - Focuses on the process by which the repository sets and manages its objectives
- Data Seal of Approval
  - Data Archiving and Networked Services (DANS), an institute of the Royal Netherlands Academy of Arts
  - *data producer; data repository; data consumer*
  - Five quality criteria (similar to RIN guidelines for stewardship of research data)

OAIS Representation Information

OAIS Representation Information (RI)

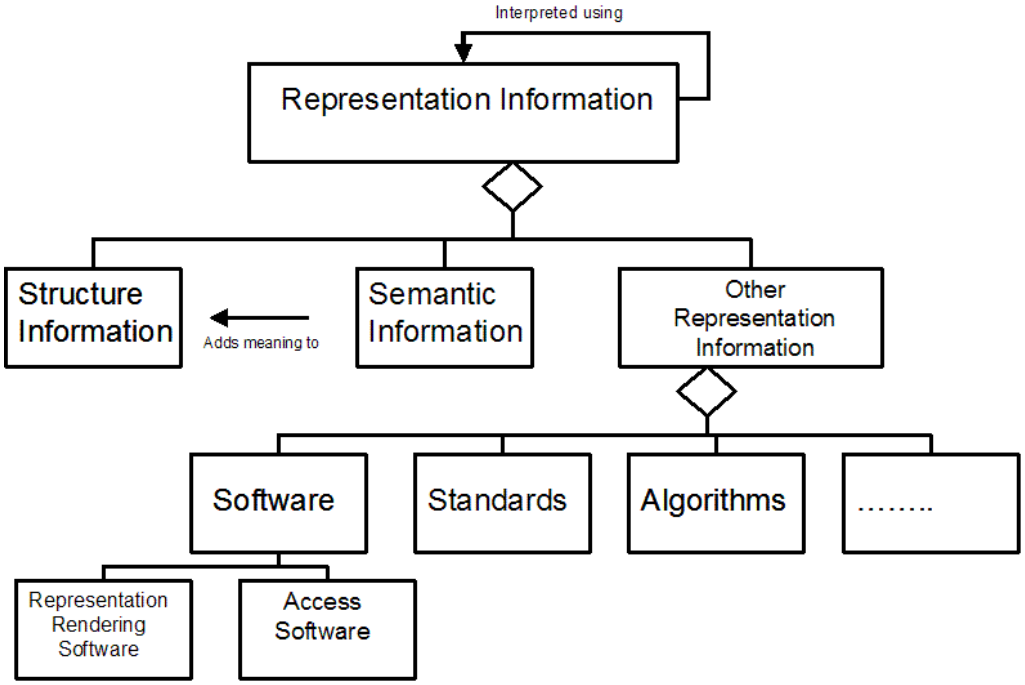
- Information Object is composed of a Data Object that is either physical or digital, as well as the Representation Information that allows for the full interpretation of the data into meaningful information
- Representation Information is *any* information required to render, interpret, process, use and understand data



OAIS Information Model
Fig 4-10
(OAIS Reference Model)

Types of OAIS RI

- Types of RI
  - Structure
    - e.g. file formats for text, images, audio, moving images, datasets, 3D models
  - Semantic
    - e.g. data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri
  - Other
    - e.g. software, algorithms, standards, time dependent information, actions, processes
- RI is recursive in nature; using one element of RI in a meaningful manner may well require further RI, resulting in a RI Network



- Recursion is terminated based on the designated community's knowledge base
- Essential that RI itself is curated and preserved to maintain access to data

Registry/Repository of RI (RRoRI)

- Development started under the DCC-Development team
- Work now being undertaken jointly with the CASPAR Project
  - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (Integrated Project co-funded by EU FP6 Programme, April 2006)
- Representation Information is the key to long-term access
- RRoRI should itself be a trustworthy OAIS
- Registry/Repository
  - Repository: some RI is stored
  - Registry: links to external RI
- Emphasis on interoperability and automated use
- Vision is to have a global, distributed network of RI
- Provide an infrastructure of reliable and trusted RI for third party use

RI Label & Curation Persistent Identifier

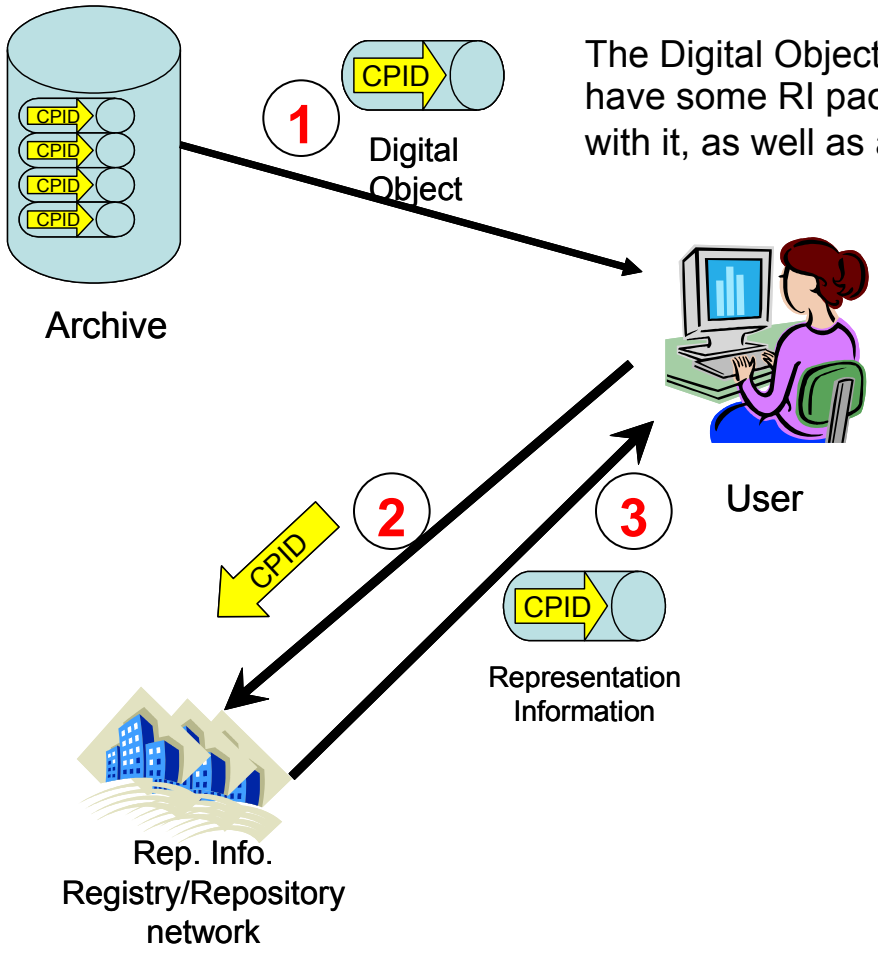
- Idea of RI is the key
  - Information Object: a specific object to be archived/preserved/ curated
  - RI: all information required to render, process, interpret, use and understand the object
  - RI Label: used to connect RI to an Information Object
- RI Label serves as a mechanism for accessing RI in RRoRI
  - Label is used to identify relevant RI
  - Provides mechanism for recording individual RI components
- RI Label has a Curation Persistent Identifier (CPID)
  - Used to connect the digital object to the RI Label

RI Network Usage Scenario

1 User gets data from archive. Data has associated Curation Persistent Identifier (CPID)

2 User unfamiliar with data so requests RI using CPID

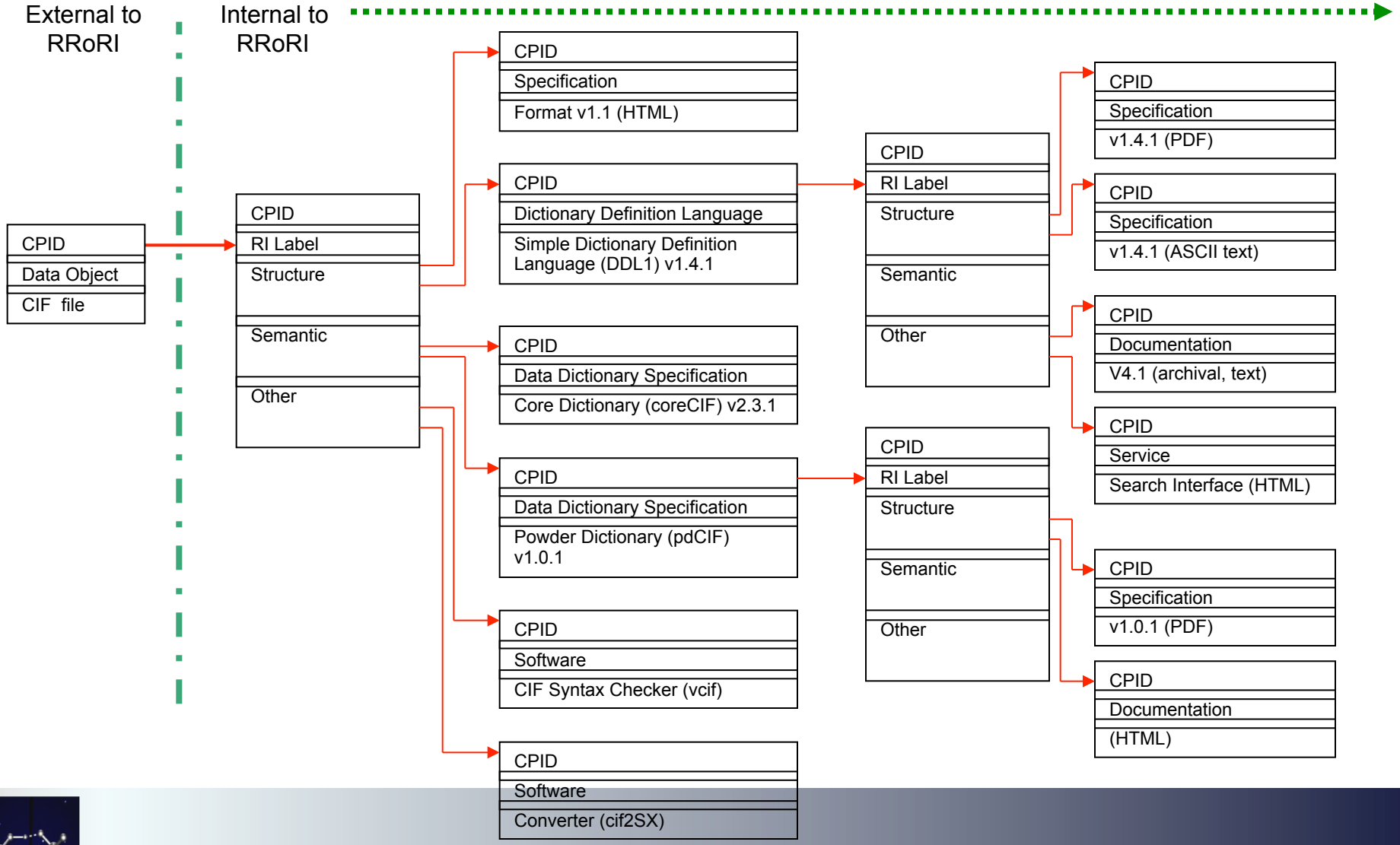
3 User receives RI – which has its own CPID in case it is not immediately usable



CPID supports automated access & processing

David Giaretta, 2007

CIF file format: Part of an RI Network



RI: Challenges and Issues

- Constructing RI Networks is time-consuming and non-trivial
  - Huge amount of information to be structured and documented
  - Take tacit, unstructured and dynamic knowledge and make it explicit with encoded relationships to enable automated processing (Semantic Web)
  - Domain expertise required for comprehensive and robust RI networks
  - Need robust search and retrieval of RI to build RI networks
- Continuous Monitoring to keep RI fit for purpose
  - Designated Community; Knowledge Base; Maintenance of RI and RI networks
- Technical Infrastructure
  - Need to record CPID as part of (preservation) metadata
  - Resolver service for CPID to enable automatic traversal of RI network
  - Continuous curation and maintenance of CPID, RI, RI Label and RI networks

Preservation Metadata

Preservation Metadata

- Documentation or metadata is fundamental to preserving and curating digital information
- Differing preservation strategies demand distinct types of information be recorded; so that the type and amount of metadata recorded will depend on the preservation strategy adopted
- OAIS Preservation Description Information (PDI)
- PREservation Metadata: Implementation Strategies (PREMIS) Data Dictionary, version 2.0, March 2008
  - Consensus on a core set of preservation metadata
  - "*things that most working preservation repositories are likely to need to know in order to support digital preservation*"

OAIS Preservation Description Information

- Reference: identifiers for unambiguous access to content

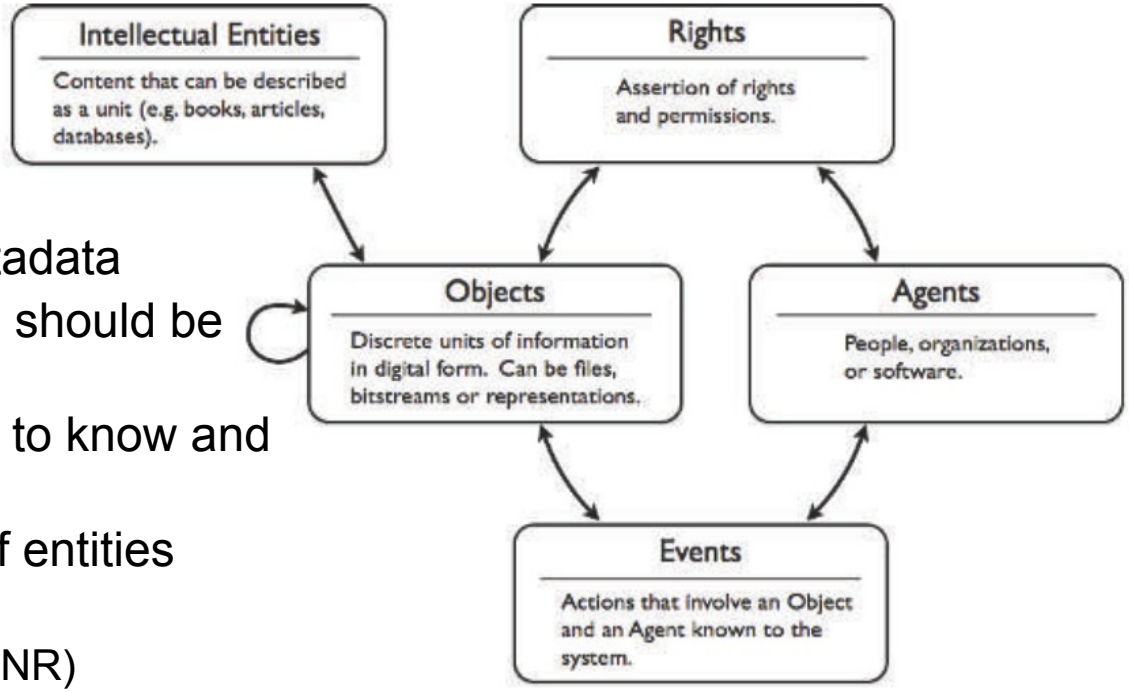
  e.g. object identifier; a journal reference; a bibliographic description or a persistent identifier.
- Provenance: history, including any changes to the data since it was submitted and who has had custody of it; provides some assurance as to the likely reliability of the data
- Context: relationship of the data to its environment and other content information.

  e.g. calibration history; relationship to other data sets; pointers to related documents etc.
- Fixity: data integrity checks to ensure that the data has not been altered in an undocumented manner; includes special encoding and error detection schemes

  e.g. checksums

PREMIS Data Dictionary V2.0

- A subset of all preservation metadata
- Does not specify how metadata should be represented in any system
- Defines what the system needs to know and should be able to export
- Semantic units are properties of entities
  - 1.1 objectIdentifier (M, R)
    - 1.1.1 objectIdentifierType (M, NR)
    - 1.1.2 objectIdentifierValue (M, NR)
  - 1.3 preservationLevel (O, R) [representation, file]
    - 1.5.4 format (M, R) [file, bitstream]



Data Model
(PREMIS Data Dictionary V 2.0)

eBank-UK Metadata Application Profile

- Simple Dublin Core
  - Crystal structure
  - Title (Systematic IUPAC Name)
  - Authors
  - Affiliation
  - Creation Date
- Qualified Dublin Core (for additional chemical metadata)
  - Empirical formula
  - International Chemical Identifier (InChI)
  - Compound Class and Keywords
- Application Profile: http://www.ukoln.ac.uk/projects/ebank-uk/schemas/
- DOI links: http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145
- Rights & Citation: http://ecrystals.chem.soton.ac.uk/rights.html

Preservation Metadata: Work in Progress

Resources:

Dataset Collection; Raw Dataset; Derived Dataset; Result Dataset; Transient Data? Workflow?

Publication/Dissemination Metadata

Persistent Identifier
Preservation Policy/Strategy
Rights management: IPR that may limit ability to preserve, disseminate or reuse the data
Context: how the data was created and under what circumstances

Management Metadata (per dataset/data file)

Embargo e.g. policy
Provenance: custodial history of the dataset
Authenticity: validation that dataset is in fact what it purports to be
Representation Information (CPID): e.g. Specifications; File formats; Software; Hardware (instrumentation); Semantics
Preservation activity: any actions taken to preserve the dataset

Conclusions

- Guaranteeing long-term safety and accessibility of fragile digital research data involves a substantial commitment
- Factors that influence preservation planning are very wide-ranging; need to be assessed in terms of their relevance to a particular context and situation
- Digital curation and preservation should be an integral part of sound data management practice
- Need digital curation throughout the useful lifetime of digital data
- Plan from the outset for longevity and sustainable access
- Cost/Benefit/Risk Analysis
  - LIFE Project
  - *Keeping Research data Safe: A cost Model and Guidance for UK Universities,* Commissioned by the JISC, April 2008
- Flexibility -plan that can be revisited and revised regularly to allow change to be managed

Questions?

Thank you

Manjula Patel
UKOLN, University of Bath, UK
m.patel@ukoln.ac.uk

eCrystals Federation Project
http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page